stagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition

Mengshi Qi, *Student Member, IEEE,* Yunhong Wang, *Senior Member, IEEE,* Jie Qin, Annan Li, *Member, IEEE,* Jiebo Luo, *Fellow, IEEE,* and Luc Van Gool

***Abstract*—In real life, group activity recognition plays a significant and fundamental role in a variety of applications, *e.g.* sports video analysis, abnormal behavior detection and intelligent surveillance. In a complex dynamic scene, a crucial yet challenging issue is how to better model the spatio-temporal contextual information and inter-person relationship. In the paper, we present a novel attentive semantic recurrent neural network (RNN), namely stagNet, for understanding group activities and individual actions in videos, by combining the spatio-temporal attention mechanism and semantic graph modeling. Specifically, a structured semantic graph is explicitly modeled to express the spatial contextual content of the whole scene, which is afterward further incorporated with the temporal factor through structural-RNN. By virtue of the 'factor sharing' and 'message passing' mechanisms, our stagNet is capable of extracting discriminative and informative spatio-temporal representations and capturing inter-person relationships. Moreover, we adopt a spatio-temporal attention model to focus on key persons/frames for improved recognition performance. Besides, a body-region attention and a global-part feature pooling strategy are devised for individual action recognition. In experiments, four widely-used public datasets are adopted for performance evaluation, and the extensive results demonstrate the superiority and effectiveness of our method.**

***Index Terms*—Group Activity Recognition, Action Recognition, Spatio-temporal Attention, RNN, Semantic Graph, Scene Understanding.**

I. Introduction

UNDERSTANDING dynamic scenes in sports games and surveillance videos encompasses a wide range of applications, like sports team tactics analysis and abnormal behavior detection. The way to recognize/understand group or cluster activities within a scene, such as 'right spiking' group activity occurred in a volleyball game [1] (see Fig. 1), is a vital yet challenging problem, because of cluttered backgrounds and mixed-up relationships, *etc*. To recognize group activity under a variety of scenarios, the contextual information ought to be taken into thought, especially the relationship or connection between actions of each person, interaction of person and objects, position of person and scene. In this paper, we concentrate on the group activity analysis and propose a novel deep model to inference the interaction between people in a scene for group activity and individual action recognition.

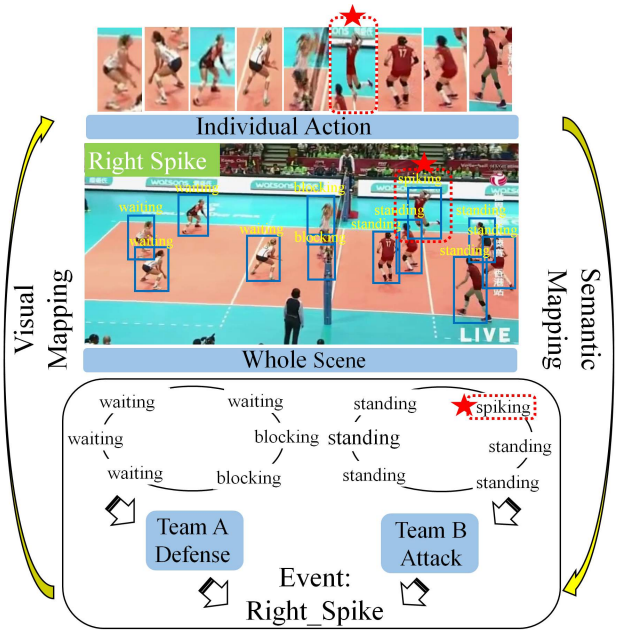


Fig. 1. Group activity understanding via a semantic graph. Using the semantic mapping, individual actions and group activity are shown on the semantic graph, which reasons inter-group relationship. Our model can also attend to the important player (with a red star) who is acting 'spiking' via the attention mechanism.

A mass of efforts [2]–[10] are devoted to handle the above issue in the computer vision community. Most of those methods try to capture spatio-temporal relations between individuals [1], [11], [12] that are considered as important cues for group activity and individual action recognition. Fundamentally, they choose the representation of visual appearance or the representation of spatial and temporal movement, for describing the dynamic interaction between people. These traditional approaches can be summarized as a mixture of hand-crafted features and probabilistic graph models. Hand-crafted features introduced in such issue consist of motion boundary histograms (MBH) [13], histogram of gradients

This work was partly supported by the National Key Research and Development Plan of China (Grant No.2016YFB1001002), the National Natural Science Foundation of China (Grant No. 61573045) and the Foundation for Innovative Research Groups through the National Natural Science Foundation of China (Grant No. 61421003). Jiebo Luo would like to thank the support of New York State through the Goergen Institute for Data Science, NSF (Awards No. 1813709 and No. 1722847), and Futurewei. Mengshi Qi acknowledges the financial support from the China Scholarship Council.

M. Qi, Y. Wang and A. Li are with Beijing Advanced Innovation Center for Big Data and Brain Computing, School of Computer Science and Engineering, Beihang University, Beijing 100191, China. E-mail: {qi_mengshi, yhwang, liannan}@buaa.edu.cn. (Yunhong Wang is the corresponding author).

J. Qin is with the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates (e-mail: qinjiebuaa@gmail.com).

J. Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627, USA. E-mail: jiebo.luo@gmail.com.

L. Van Gool is with the Computer Vision Laboratory, ETH Zurich, 8092 Zurich, Switzerland. E-mail: vangool@vision.ee.ethz.ch.

(HOG) [14], the cardinality kernel [15], *etc*. Markov Random Fields (MRFs) [16] and Conditional Random Fields (CRFs) [17] also have been utilized to simulate the inter-object relationships.

An obvious limitation of the above-mentioned approaches is that the low-level features fall short of representing advanced group activities. With the success of convolutional neural networks (ConvNets) [18]–[20] in several computer vision tasks, deep feature representations have demonstrated their capabilities in representing advanced visual appearance. However, typical ConvNets regard single frame of a video as input and output a holistic feature vector by average pooling, hence spatial and temporal relations between consecutive frames cannot be discerned. The spatio-temporal relations [1], [11], [12] comprise the spatial appearance and temporal action of each person and their interaction. Recurrent Neural Networks (RNNs) [21], [22] have the power to represent dynamic temporal actions from the sequential data with the temporal features. Therefore, it is extremely fascinating to explore an RNN based network for capturing the crucial spatio-temporal contextual information.

Moreover, automatically describing the semantic contents within the scene is rewarding to better understanding the overall hierarchical structure of the scene (*e.g.* sports games and surveillance videos). However, this task is extremely troublesome, because the semantic description not only captures the individual action, but also completely expresses how these persons relate to each other and how the whole group event happens. If the above RNN based network also can describe the semantic contents of the dynamic scene, we can have a substantially much clearer understanding.

In this paper, we present a novel attentive semantic recurrent neural network, called *stagNet* for group activity and personal action recognition, which combines spatial-temporal attention and semantic graph. Specifically, individual activities and their spatial relations are inferred and depicted by an explicit semantic graph, and their temporal interactions are integrated by a structural-RNN model. We utilize "message passing" mechanism to transport spatial and temporal semantic information between different components (*e.g.* nodeRNN and edgeRNN), and adopt "factor sharing" mechanism to permit the equivalent element in temporal dimension shares the identical spatial factor. The model passes the message that incorporates contextual semantic features between every element of the graph. Furthermore, a spatio-temporal attention mechanism is incorporated into for leveraging various levels of importance to different persons/body-regions/frames in video sequences. More significantly, the semantic graph and spatio-temporal attention are collaboratively end-to-end trained. Lastly, we have done extensive experiments in four datasets, i.e. Collective Activity Dataset [23], New Collective Activity Dataset [24], UCLA Courtyard Dataset [7], and Volleyball Dataset [1], and the performance of our framework demonstrates that our stagNet is capable to model complex and advanced relationship, and recognize group activity and personal action.

It oughts to be mentioned that this paper is an extended version of our previous conference paper [25]. Compared to the preliminary version, we present an individual body-region attention mechanism and a global-part feature pooling strategy for improved the performance of individual action recognition. Moreover, during the testing process, we conduct extensive experiments on two more public benchmark datasets, *i.e. New Collective Activity Dataset* [26] and *UCLA Courtyard Dataset* [7], perform more qualitative results and detailed analysis to demonstrate the effectiveness of our proposed stagNet in this journal paper.

Precisely, the main contributions of this paper include:

- We introduce a semantic graph to describe explicitly all the content in the scene, *i.e.* group activity, individuals' actions and their spatial relations, with a 'message passing' mechanism. To the best of our knowledge, we are the first to output such a semantic graph for understanding group activities.
- We extend our semantic graph model to the temporal dimension between frames in a video via a structural-RNN, which is achieved by adopting the 'factor sharing' mechanism.
- A spatio-temporal attention mechanism and global-part feature pooling operation are further integrated for better performance, which places stress on the most representative persons, the vital body region of individual player or the crucial frames within the video.
- Experimental results on four public benchmark datasets demonstrate that the performance of our framework is competitive with that of the state-of-the-art approaches.

The rest of our paper is organized as the following. Section II reviews related works on group activity recognition and deep structure model in brief. In Section III, we elaborate the proposed method in detail. Afterward, Section IV presents the experimental results and comprehensive analysis at length. Finally, we draw the conclusion of this work in Section V.

II. Related Work

In this section, we review the related works concisely. We firstly introduce group activity recognition, which is the most relevant works to this paper. Then, we survey recent advances in modeling structures and attention mechanisms in deep learning.

Group Activity Recognition. Traditional approaches usually capture hand-crafted features as spatio-temporal representations (*e.g.*, MBH and HOG) [23]. Khamis *et al.* [28] combined per-frame and per-track cues for action recognition in a structured scene. A multi-agent event detection method was presented [29] with quadratic programming and linear programming for role and event localization. Considering the individual actions and interactions between different persons in a scene, several graph models were introduced to handle this problem [3]–[6], [30], [31]. Lan *et al.* [3] proposed an adaptive structure modeling algorithm to model the latent time-space structure. Wang *et al.* [4] learned a Markov Random Fields (MRFs) [32] graph to model the complicated dependencies in human cluster activity and individual action. Amer *et al.* [5] conducted a Hierarchical Random Field (HiRF) to extract the interaction between grouping nodes and the hidden variables

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2894161, IEEE Transactions on Circuits and Systems for Video Technology

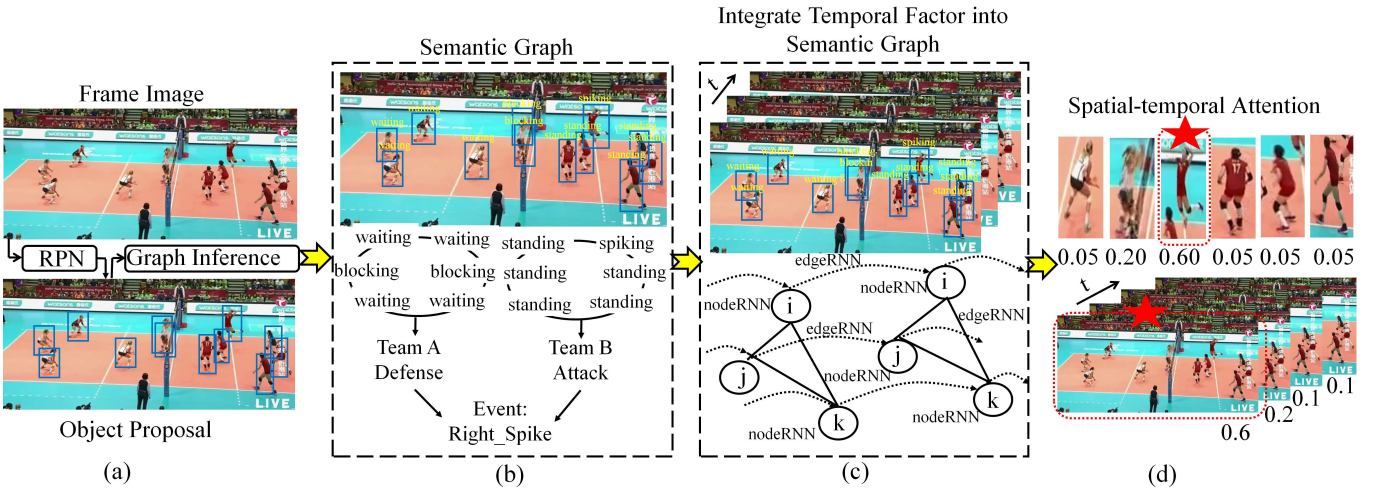


Fig. 2. Pipeline of our proposed stagNet. From left to right: (a) object proposals are obtained from raw frames via a region proposal network (RPN) [27]; (b) the semantic graph is modeled and constructed from text labels and visual data; (c) temporal factor is integrated into the graph by leveraging a structural-RNN, and the semantic graph is inferred by message passing and factor sharing mechanisms; (d) finally, a spatio-temporal attention mechanism is introducing for placing emphasis key persons/body regions/frames (denoted with a red star) for better performance.

in a scene. Shu *et al.* [6] formulated a spatio-temporal AND-OR graph for interaction inference between groups or teams, events and human roles jointly [7]. However, these approaches utilized shallow features that cannot encode advanced semantic information, and frequently overlook temporal relationship.

Recently, a number of deep models [1], [11], [25], [33]–[38] have been devised for the task of group activity recognition. As an example, Deng *et al.* [11] presented a graphical model with a gate function to simulate edges and nodes. Ibrahim *et al.* [1] proposed a hierarchical deep model, which contains first LSTM layer to extract individual person's dynamic action, and the second LSTM layer to recognize group activity. Shu *et al.* [33] introduced confidence-energy recurrent networks (CERN) with a novel energy layer, which can compute the p-values to estimate the prediction energy for recognizing group activity. Wang *et al.* [34] introduced a recurrent interaction framework, which unified all the contextual features, *e.g.* individual person, inter-group and intra-group interactions. In [35], multi-class object detection [19], [39] and fully convolutional network [40] were adopted to capture multi-scale features for estimating individual actions and collective activities with probability inference. Afterward, Li *et al.* [36], [41] introduced an image captioning and optical flow-based model, and Biswas *et al.* [37] adopted a series of structural interconnected RNNs to address the problem. Ibrahim *et al.* [38] proposed a hierarchical relational network for group activity recognition and retrieval. Moreover, Fan *et al.* [42] proposed a multi-instance learning method for complex event detection by adaptively selecting reliable shots. However, most of these works either extracted individual features in spite of the scene context or captured the context in an implicit manner without any semantic information. In this paper, we conceive to *explicitly* capture the semantic context of the scene by an expressive spatio-temporal *semantic* graph [43] through RNNs.

Deep Structure Model. A plenty of researches have been devoted to forming a more powerful deep network by combining graph models. Bengio *et al.* [16] adopted ConvNets integrated with Hidden Markov Model [44] for handwriting recognition. Chen *et al.* [45] incorporated Markov Random Fields (MRFs) into a deep learning model, and then Liu *et al.* [46] adopted a similar architecture for semantic segmentation. [47]–[49] introduced deep neural networks with graph-structure learning for estimating human pose. "DeepLab" system was performed in [50] for image segmentation, which combined deep ConvNets with fully-connected conditional random fields (CRF) [17]. Zheng *et al.* [51] integrated CRF-based probabilistic graphic model with the recurrent neural network for semantic segmentation. Zhang *et al.* [52] introduced the Bayesian optimization with ConvNets [53] for improving object detection. Defferrard *et al.* [54] designed fast localized convolutional filters on graph-based ConvNets within the context of spectral graph theory. Niepert *et al.* [55] introduced a framework for learning convolutional neural networks for arbitrary graphs, which contains discrete and continuous node and edge attributes. However, most of the aforementioned works are task-specific and probably fail to model spatio-temporal and interaction information from dynamic videos. Structural-RNN [12] was a Recurrent Neural Networks combined with the advanced spatio-temporal graph structure. Inspired by [12], we explicitly construct a semantic graph under a spatio-temporal manner and describe semantic space-time contents of the scene, *e.g.* inter-object and intra-person relationships.

Attention Mechanism. Attention mechanisms have been widely applied in vision and language tasks and achieved great success. The pioneering research [56] proposed the visual saliency detection as an attention model for scene recognition. Shapobalova *et al.* [57] utilized human eye gaze as attention to recognize actions in video. Afterward, Mnih *et al.* [58] firstly incorporated attention strategy into RNNs to extract selected regions in order. The "Look and think twice" mechanism proposed by [59] was able to capture visual attention on specific objects in images with deep learning model. An active object detection based on dynamic attention-action strategy was introduced in [60]. Attention models have been additionally applied in machine translation [61] and

image captioning [62] with natural language processing. Xu *et al.* [62] devised soft attention and hard attention for image caption. A temporal attention mechanism with text-generation RNNs was introduced in [63] to choose the foremost related frames. Ramanathan *et al.* [64] adopted time-varying attention feature learned from the recurrent neural network (RNN) for key player tracking and event classification. Ba *et al.* [65] introduced "fast weight" that can be used to store and attend to the temporary memories of the recent past. In our proposed stagNet, we incorporate the contextual semantic graph and spatial-temporal attention into a unified framework, which is trained to focus on more relevant persons, individual body regions and frames jointly in the video.

III. The Proposed Approach

The overall architecture of proposed stagNet for group activity and personal action recognition is depicted in Fig. 2 and Fig. 3. We employ hierarchical RNN with two varieties of RNN units (*i.e.* nodeRNN and edgeRNN) and train them in an end-to-end manner. Above all, the first step is to construct the semantic graph using each frame as input, and then we incorporate the temporal factor by employing a structural RNN. The inference is implemented by virtue of 'message-passing' and 'factor sharing' strategy. Finally, we introduced a spatio-temporal attention mechanism to discover crucial people, body regions and frames for performance improvement.

A. Semantic Graph

In this subsection, we introduce the semantic graph and the mapping from visual data to the graph. We inference the semantic graph to predict person's affiliations based on their positions and visual appearance. As shown in Fig. 2(b), the semantic graph is constructed by parsing a scene with multiple people into a collection of bounding boxes related to the corresponding spatial positions. Each bounding box of a specific person is defined as a node of the graph. The graph edge that describes pairwise relations is determined by the spatial distance and temporal correlation, which will be introduced in Section III-B.

To generate a collection of person-level proposals (bounding boxes) from the t-th frame I^t in video I, we utilize the region proposal network (RPN), which is an element of the region-based fully convolutional networks [27]. The RPN outputs position-sensitive score maps as the relative position, and connects a position-sensitive region-of-interest (RoI) pooling layer on top of the fully convolutional layer. These proposals are considered as the input of the graph inference procedure. A total of three kinds of information are inferred by modeling the graph: (1) the individual action label, (2) the inter-group relationships, and (3) the group activity label.

Within the t-th frame I^t, we define a collection of K bounding boxes as $B_{I^t} = (x_{t,1}, ..., x_{t,K})$, and the inter-person relationship set as R (*e.g.* whether two players belong to the identical team). Given a set of the scene labels (*i.e.* group activity label) C_{scene}, and individual action labels set C_{action}, we define $y^t \in C_{scene}$ as the scene category label, $x_i^{act} \in C_{action}$ as the action category label of the i-th person proposal, x_i^{pos} as its spatial coordinates, and $x_{i\to j} \in R$ as the predicted relationship between the i-th and j-th person proposal boxes. Besides, we define the set of all variables to be $x = \{x_i^{act}, x_i^{pos}, x_{i\to j} \mid i = 1...K, j = 1...K, j \neq i\}$. In particular, the semantic graph is constructed by seeking out the optimal y^{t*} and x^* that maximize probability function as follows:

$$
\begin{aligned}
&< x^*, y^{t*} >= arg \max_{x, y^t} Pr(x, y^t \mid I^t, B_{I^t}), \\
&Pr(x, y^t \mid I^t, B_{I^t}) = \prod_{i,j \in K} \prod_{j \neq i} Pr(y^t, x_i^{act}, x_i^{pos}, x_{i\to j} \mid I^t, B_{I^t}).
\end{aligned} \tag{1}
$$

In the following, we are going to present the graph inference of frame-wise semantic graph structure in detail.

B. Graph Inference

Inspired by [66], the graph inference is applied by adopting the mean field and computing the hidden states via Long short-term memory (LSTM) network [21], which could be a simplified yet effective recurrent neural network. We define the semantic graph as $G = (S, V, E)$, where S is the scene node, and V and E are the object nodes and edges respectively. Concretely, S represents the global representation of a frame in a video, an object node $v_i \in V$ $(i = 1, \cdots, K)$ refers to the person-level proposal (where $i = 1, ..., K$ corresponds to the totally K persons in the scene), and the edge E corresponds to the spatial configuration of object nodes V in the frame. Through the mean field inference, we approximate $Pr(x, y^t \mid \cdot)$ by $Q(x, y^t \mid \cdot)$ that depends on the current state of each node and edge. The hidden state of LSTM unit is the current state of each node and edge in the semantic graph. We define h^t as the current hidden state of scene node, h_{v_i} and $h_{e_{ij}}$ as the current hidden state of node i and edge $i \to j$ respectively. Note that all the nodeRNNs share the identical set of parameters and all the edgeRNNs share another set of parameters. The solution to $Q(x, y^t \mid I^t, B_{I^t})$ will be achieved by calculating the mean field distribution as the following:

$$
\begin{aligned}
&Q(x, y^t \mid I^t, B_{I^t}) \\
=&\prod_{i=1}^{K} Q(x_i^{act}, x_i^{pos}, y^t \mid h_{v_i}, h^t) Q(h_i \mid f_{v_i}) Q(h^t \mid f^t) \\
&\prod_{j \neq i} Q(x_{i\to j} \mid h_{e_{ij}}) Q(h_{e_{ij}} \mid f_{e_{ij}}),
\end{aligned} \tag{2}
$$

where f^t is the convolutional feature of the scene in the t-th frame, f_{v_i} is the feature of the i-th node, and $f_{e_{ij}}$ is the feature of the edge connecting the i-th node and j-th node, which is the unified bounding box over two nodes. We compute $f_{e_{ij}}$ using six features via calculating the basic distances and direction vectors, *i.e.* $< |dx|, |dy|, |dx + dy|, \sqrt{(dx)^2 + (dy)^2}, arctan(dy, dx), arctan2(dy, dx) >$. All of these features are captured by the RoI pooling layer. Then the messages aggregated from other previous LSTM units are input to the next step.

As shown in Fig. 2, the edgeRNNs offer contextual information for the nodeRNNs, and the max pooling is performed over the nodeRNNs. The nodeRNN concatenates the node feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2894161, IEEE Transactions on Circuits and Systems for Video Technology

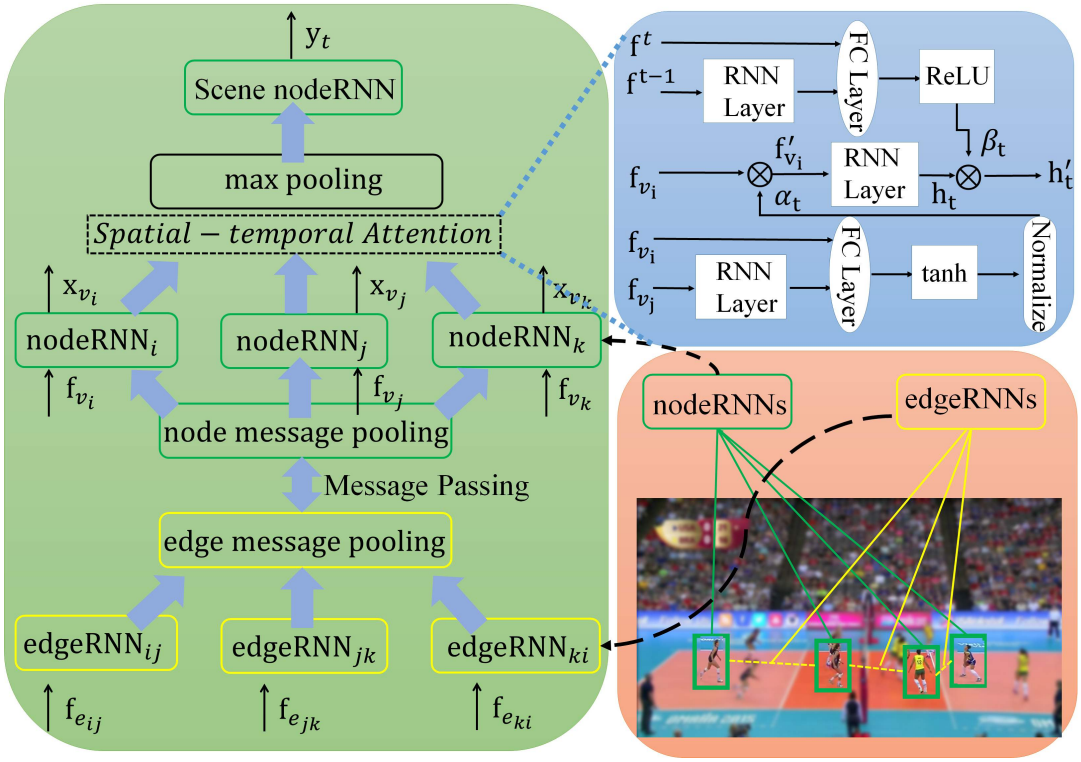


Fig. 3. Illustration of our nodeRNN and edgeRNN model. At first, the model extracts visual features of nodes and edges from a group of object proposals, and then takes the visual features as initial input to the nodeRNNs and edgeRNNs. We introduce the node/edge message pooling to update the hidden states of nodeRNNs and edgeRNNs. The input of nodeRNNs is the output of the edgeRNNs, and nodeRNN also output the label of personal action, and the max pooling is performed subsequently. Furthermore, a spatio-temporal attention mechanism is incorporated into our architecture. Finally, the top-most nodeRNN (*i.e.* Scene nodeRNN) outputs the label of group activity.

and the outputs of edge-RNN accordingly. And the edgeRNN passes the summation of all edge features that are connected to the identical node as the message. EdgeRNNs and nodeRNNs take the visual features as initial input and produce a collection of hidden states. The model iteratively updates the hidden states of the RNN. Finally, the hidden states of the RNN are utilized to predict frame-wise scene label, person action label, person position information and inter-group relationships.

Message passing [66] is able to iteratively improve the effectiveness of inference in the semantic graph. In the graph topology, the neighbors of the egdeRNNs are nodeRNNs. Passing messages through the entire graph involves two sub-graphs: *i.e.* node-centric sub-graph and edge-centric sub-graph respectively. For node-centric sub-graph, the nodeRNN receives messages from its neighboring edgeRNNs. Similarly, for edge-centric sub-graph, the edgeRNN gets messages from its adjacent nodeRNNs. We introduce an aggregation function referred to as message pooling to learn accommodative weights for modeling the importance of passed messages. We calculate the weight factors for each incoming message and aggregate the messages by a total weight for final representation. It demonstrates that this strategy is more effective than average-pooling or max-pooling [66].

Specifically, we denote the update message input to the i-th node v_i as m_{v_i}, and message to the edge between the i-th and j-th node e_{ij} as $m_{e_{ij}}$ respectively. Then, we calculate the message passed into the node considering its own hidden state h_{v_i} and the hidden state of its connected edges $h_{e_{ij}}$ and $h_{e_{ji}}$, and acquire the message passed into edge with respect to the hidden state of its adjacent nodes h_{v_i} and h_{v_j}. Formally, m_{v_i} and $m_{e_{ij}}$ are computed as

$$
\begin{aligned}
m_{v_i} &= \sum_{j:i\to j} \sigma(U_1^T[h_{v_i}, h_{e_{ij}}])h_{e_{ij}} + \sum_{j:j\to i} \sigma(U_2^T[h_{v_i}, h_{e_{ji}}])h_{e_{ji}}, \\
m_{e_{ij}} &= \sigma(W_1^T[h_{v_i}, h_{e_{ji}}])h_{v_i} + \sigma(W_2^T[h_{v_j}, h_{e_{ij}}])h_{v_j},
\end{aligned} \tag{3}
$$

where W_1, W_2, U_1 and U_2 are parameters to be learned, σ is defined as a sigmoid function, and $[\cdot,\cdot]$ means concatenation of two hidden vectors. Finally, we utilize these messages to update the hidden state of nodeRNN and edgeRNN iteratively. Once finishing updating, the hidden states are then used to predict personal action classes, bounding box offsets and relationship varieties.

C. Integrating Temporal Factors

With the semantic graph of a frame, temporal factors are integrated to construct the spatio-temporal semantic graph (see Fig. 2(c)) with the structural-RNN [67]. Based on the graph definition in Section III-A and III-B, we add a temporal edge E_T, such that $G = (S, V, E_S, E_T)$, where E_S denotes the spatial edge. The node $v_i \in V$ and edge $e \in E_S \cup E_T$ in the spatio-temporal semantic graph enrolls over time. In particular, the nodes at neighbor time steps, *e.g.* the node v_i at time t and the node v_i at time $t+1$ are connected with the temporal edge $e_{ii} \in E_T$. We define the node label as y_v^t and corresponding feature vectors for node and edge are referred to f_v^t, f_e^t at time t, respectively. We present a 'factor

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2894161, IEEE Transactions on Circuits and Systems for Video Technology

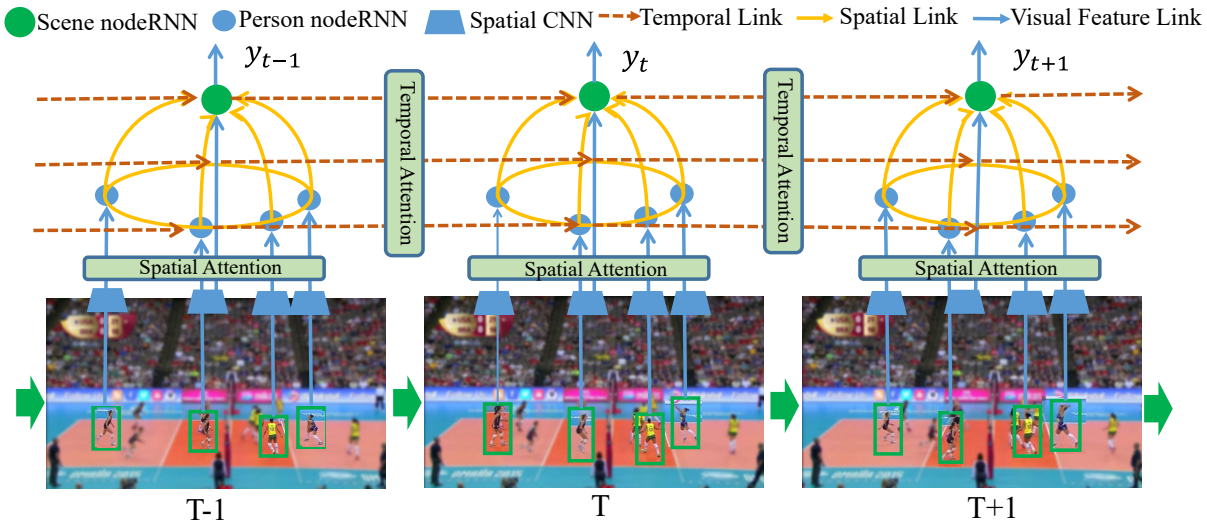


Fig. 4. Hierarchical semantic RNN structure for a volleyball match. Given object proposals and tracklets of all players, we feed them into spatial ConvNet, followed by an RNN to represent each player's action and appearance of the full scene. Then we adopt structural-RNN to determine temporal links for a sequence of frames. Furthermore, we integrate the LSTM based spatio-temporal attention mechanism into the model. The output layer classifies the entire team's group activity.

sharing' mechanism, which indicates that the nodes denoting the identical person and the edges representing the same relationship tend to share factors (*e.g.* parameters, original hidden states of RNNs) across different video frames. Fig. 4 illustrates an example of structural-RNN through three time steps in a volleyball match video. Please refer to [67] for more technical details concerning structural-RNN.

We define two varieties of edges (edgeRNN) in the spatio-temporal graph: One is spatial-edgeRNN representing the spatial relationship. It is formed by the spatial message pooling function in each frame and computed from adjacent player's nodeRNN depended on Euclidean distance. The other is temporal-edgeRNN that connects adjacent frames of the same player to represent the temporal information, which is created by sharing factors between players' nodeRNNs in a sequence of video. We incorporate the features of spatial edgeRNN between two consecutive frames into temporal edgeRNN, leading to twelve additional features.

Throughout training, the errors of predicting the labels of scene nodes and object nodes are back-propagated through the sceneRNN, nodeRNN and edgeRNN. The passed messages represent the interactions between nodeRNNs and edgeRNNs. The nodeRNN is connected to the edgeRNN, and outputs the personal action labels. Each edgeRNN models the semantic interaction between adjacent nodes simultaneously and the evolution of interaction over time.

D. Spatial-Temporal Attention Mechanism

The group activity involves multiple persons, however only few of them play significant roles in determining the activity. For instance, the 'winning point' activity in a volleyball game typically occurs with a specific player spiking the ball and another player failing to catch the ball. For a much better understanding of the group activity and individual action, it is rewarding to assign higher levels of importance to decisive persons and critical body regions of the individual player. Inspired by [64], [68], we introduce a spatio-temporal soft attention mechanism to focus on specific persons and individual body components in specific frames to enhance the recognition accuracy of group activity and individual action, as illustrated in Fig. 4. It is worthy noted that we combine proposals of the identical person with KLT trackers [69] to construct the complete representation of a player information from a sequence of frames. While the person detections vary from one frame to another, they can be associated across frames through tracking, which leads to better feature representation of the players.

1) Person-Level Spatial Attention: We adopt a spatial attention model to set weights to different persons using long-short term memory (LSTM) networks. Concretely, given one frame that has K players $x_t = (x_{t,1}, ..., x_{t,K})$, we define the scores $s_t = (s_{t,1}, ..., s_{t,K})^T$ as the importance of all person actions in each frame:

$$s_t = W_s \tanh(W_{xs} x_t + U_{hs} h^s_{t-1} + b_s), \tag{4}$$

where W_s, W_{xs}, U_{hs} are the learnable parameter matrices, and b_s is the bias vector. h^s_{t-1} is the hidden variable from an LSTM unit. And for the k-th person, the spatial attention weight is calculated as a normalization of the scores:

$$\alpha_{t,k} = \frac{\exp(s_{t,k})}{\sum_{i=1}^{K} \exp(s_{t,i})}. \tag{5}$$

Afterward, the input to LSTM unit is updated as $x'_t = (x'_{t,1}, ..., x'_{t,K})^T$, where $x'_{t,k} = \alpha_{t,k} \cdot x_{t,k}$. Then the representation of the attended player can be regarded as the input to the RNN nodes in the spatio-temporal semantic graph mentioned in Section III-A.

2) Individual Body-Region Attention: For individual action recognition, different body regions contribute to the ultimate result in variational weights. As an example, the movement of arms is more critical than other body components for 'spiking' action of a player. It is indispensable to seek out a way for understanding individual posture and limb articulation. Hence, we propose a soft attention to attend different body regions,

for generating an attentive individual spatial feature $x_{t,K}^{att}$ to replace $x_{t,K}$ in above-mentioned, and performing a global-part pooling strategy both considering global visual features $x_{t,K}^{gl}$ and body-region visual features $x_{t,K}^{body}$. For the K-th player in the t-th frame, we denote the individual global visual feature as $x_{t,K}^{gl}$, and we divide each person into six parts equally (*i.e.* three rows and two columns w.r.t. right/left head, right/left upper body and right/left bottom body, as portrayed in Fig. 5) that refers to $x_{t,K_{br}}$, $br \in \{1:6\}$. Then the vital scores $\gamma_{br} = (\gamma_1, \cdots, \gamma_6)^T$ of each body-region are computed as the following:

$$\gamma_{br} = softmax(W_\gamma(\tanh(W_{x\gamma}x_{t,K_{br}} + U_{h\gamma}h_{t-1}^{\gamma}))), \quad (6)$$

where W_γ, $W_{x\gamma}$ and $U_{h\gamma}$ are the parameters to be learned, h_{t-1}^{γ} is the hidden vector in an LSTM unit. Then, the attentive appearance representation of individual player is denoted as $x_{t,K}^{att} = [x_{t,K}^{gl}, x_{t,K}^{body}]$, where $x_{t,K}^{body} = \sum_{br=1}^{6} \gamma_{br} x_{t,K_{br}}$, and $[\cdot,\cdot]$ refers to the concatenation operation as global-part pooling. Note that we only adopt individual body-region attention with given ground-truth personal bounding box for individual action recognition instead of group activity recognition. The visualization of our person-level spatial attention and individual body-region attention models are illustrated in Fig. 5.

3) Frame-Level Temporal Attention: In a sequence, only a number of frames contain the vital information. In order to find them, we apply a temporal attention model to assign weight β to each frame. For T frames in a video, the temporal attention model is comprised of an LSTM layer, a fully connected layer and a nonlinear ReLU unit. The temporal attention weight of the t-th frame can be calculated as

$$\beta_t = ReLU(W_{x\beta}x_t + U_{h\beta}h_{t-1}^{\beta} + b_\beta), \quad (7)$$

where x_t is the current input and h_{t-1}^{β} is the hidden variables at time step t-1. The temporal attention weight controls how much information of every frame can be used for making the final recognition decision. Obtaining the output z_t of the main LSTM network and the temporal attention weight β_t at each time step t, the important scores for C_{scene} classes are the weighted summation w.r.t at all time steps:

$$o = \sum_{t=1}^{T} \beta_t \cdot z_t, \quad (8)$$

where $o = (o_1, o_2, \cdots, o_{C_{scene}})^T$, and T indicates the number of frames. The probability of being the i-th category for video I is

$$p(C_{scene}^i|I) = \frac{e^{o_i}}{\sum_{j=1}^{C_{scene}} e^{o_j}}. \quad (9)$$

E. Joint Objective Function

Finally, we formulate the joint overall objective function with a regularized cross-entropy loss, and combine the semantic graph modeling and the spatio-temporal attention network learning as the following:

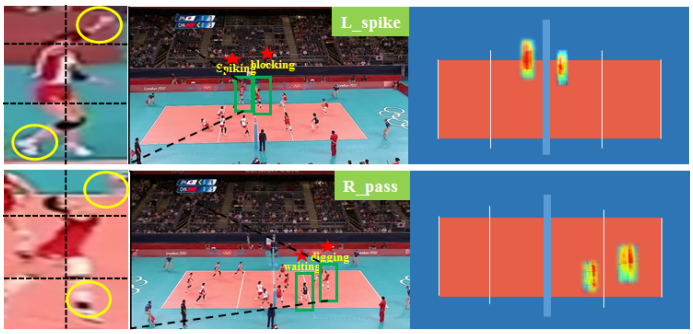


Fig. 5. Visualization of our person-level spatial attention and individual body-part attention model. We visualize the distribution of attention to different people in terms of two group activities. The top row shows the players who are 'spiking' and 'blocking', which are important for the 'Left spike' activity. The bottom row shows the players who are 'digging' and 'waiting', which are important for the 'Right pass' activity. Besides, the left column shows the critical body movements to determine individual action, *i.e.* arms and legs, of two key players.

$$\begin{aligned} L = &- \sum_{i=1}^{C_{scene}} y^i \log \hat{y^i} - \frac{1}{K}\sum_{i=1}^{K} x_i^* \log \hat{x_i^*} + \\ &\lambda_1 \sum_{k=1}^{K}(1 - \frac{\sum_{t=1}^{T}\alpha_{t,k}}{T})^2 + \frac{\lambda_2}{T}\sum_{t=1}^{T}\|\beta_t\|_2 + \lambda_3\|W\|_1, \end{aligned} \quad (10)$$

where y^i and x_i^* refer to the ground-truth category label of group activity and personal action, respectively. If a video sequence is classified as the i-th class, $y^i = 1$ and $y^j = 0$ for $j \neq i$. $\hat{y^i} = p(C_{scene}^i|I)$ is the probability that a sequence is classified as the i-th class. $\hat{x_i^*} = p(C_{action}^i|B_{I^t})$ is the probability that a personal action belongs to the i-th class. For recognition, we employ max-pooling over the hidden representations followed by a softmax classifier. λ_1, λ_2 and λ_3 denote regularization terms. The third regularization term is introduced to ensure to attend to more persons within the spatial scene, and the forth term regularizes the learned temporal attention weights via l_2 normalization. The last term regularizes all the parameters of the spatio-temporal attention mechanism [68].

IV. Experiments

In this section, we extensively evaluate the performance of our stagNet on four public benchmark datasets, *i.e.* Collective Activity Dataset [23], New Collective Activity Dataset [26], UCLA Courtyard Dataset [7] and Volleyball Dataset [1] in terms of two tasks, *i.e.* group activity and personal action recognition. In the following, we first introduce the datasets and implementation details in brief. Then we describe the compared baselines, and present the experiments results and corresponding analysis.

A. Datasets

Collective Activity Dataset [23] consists of 44 video clips in total (about 2,500 frames shot by low-resolution cameras), five group activities: *crossing, waiting, queuing, walking* and *talking*, and six individual actions: *N/A, crossing, waiting,*

queuing, walking and *talking*. The category label of group activity is determined depending on the majority of people's actions classed in a video clip. The representations of the entire scene are modeled as a bag of features descriptors of individual action. Following the equivalent experimental setting in [3], we adopt directly the tracklet data released in [26], and choose $1/3$ of the video clips for testing and the rest for training.

New Collective Activity Dataset [26] totally contains 32 videos showing six collective activities: *gathering, talking, dismissal, walking together, chasing* and *queuing*. Also, it includes nine interactions: *approaching, walking-in-opposite-direction, facing-each-other, standing-in-a-row, walking-side-by-side, walking-one-after-the-other, running-side-by-side, running-one-after-the-other* and *no-interaction*, three individual actions: *walking, standing still* and *running*. Based on [11], we adopt the setting with 2,241 frames for training and 1,106 for testing, as a result of leave-one-out for testing is not improper for deep learning methods. Meanwhile, we employ the trajectory data provided in [26].

UCLA Courtyard Dataset [7] contains a 106-minute, 30 fps, 2560×1920-resolution video footage totally, which shows two distinct scenes from a bird-eye viewpoint of a courtyard at the UCLA campus. There are totally six group activities (*i.e. walking-together, standing-in-line, discussing-in-group, sitting-together, waiting-in-group* and *guided-tour*), and ten primitive actions (*i.e. riding-skateboard, riding-bike, riding-scooter, driving-car, walking, talking, waiting, reading, eating* and *sitting*) are annotated. Following [7], we split the dataset 50%/50% for training and testing.

Volleyball Dataset [1] consists of 55 video clips of volleyball games with 4,830 labeled frames in total. Each player is annotated with a bounding box and one amongst nine individual action labels: *waiting, setting, digging, falling, spiking, blocking, jumping, moving* and *standing*. Each full frame is labeled with one amongst eight group activity categories: *right set, right spike, right pass, right winpoint, left winpoint, left pass, left spike* and *left set*. Following the equivalent setting in [1], we select $2/3$ of the dataset as training set and the rest $1/3$ as the testing set. Especially, we divide all players in each frame into two teams following [1], and define four team-level activities additionally: *attack, defense, win* and *lose* in our experiments.

B. Implementation Details

Our model is implemented based on the TensorFlow [72] framework. We select the VGG-16 model [70] pre-trained on ImageNet as the backbone for extracting dense features, and only adopt the convolution layers of VGG-16 and concatenate a 1024-d 1×1 convolutional layer. Then, each frame is represented by a 1024-d feature vector. Based on the RPN detector [27], each bounding box is represented as a 2805-d feature vector, which contains 1365-d appearance representation and 1440-d spatial representation. Specifically, the appearance features can be extracted by feeding the cropped and resized bounding box through the backbone network, and utilizing spatially pooling to achieve the response map from a lower layer. The normalization operations include rescaling distance features based on width, height, and area size of frames, and resizing the summation of edge features based on the number of edges associated with each node. Furthermore, we directly adopt the tracklet data released on each dataset for the ground truth person bounding boxes, *i.e.* Collective Activity/New Collective Activity/UCLA Courtyard/Volleyball, and utilize the KLT trackers [69] to acquire tracklet for each person proposal.

The LSTM layers in our stagNet used as nodeRNN and edgeRNN have 1024-d hidden units, and they are trained by adding a softmax loss on the top at each time step. The softmax layer is utilized to produce the score maps for the action category and group activity class. Then, we additionally add a fully connected layer for regressing the offset of each person bounding box. We adopt a sliding window of 10 frames, and the batch size is set based on the analysis of experimental results. To obtain the best performance, the batch size for training the bottom layer of LSTM and the fully connected layer of RPN is 8, and the training is performed within 20,000 iterations. The top layer of LSTM is trained in 10,000 iterations with a batch size of 32. Moreover, the number of nodeRNNs in our model corresponds to the number of detected persons, which is adjustive for various persons in different events on each dataset. Besides, all the nodeRNN and edgeRNN (*i.e.* spatial-edgeRNN and temporal-edgeRNN) in our proposed stagNet share weights, respectively, according to the "factor sharing" mechanism. "Factor sharing" can leverage the nodeRNNs and edgeRNNs in temporal dimension to share the same spatial information as factor. For optimization, we employ RMSprop [73] with a learning rate ranging from 0.00001 to 0.001 for mini-batch gradient descent. Practically, we set $\{\lambda_1,\lambda_2,\lambda_3\}$ as $\{0.001,0.0001,0.0001\}$ for Collective Activity/New Collective Activity/UCLA Courtyard, and $\{0.01,0.001,0.00001\}$ for Volleyball. In addition, the semantic graph outputted of proposed stagNet is recorded as a JavaScript Object Notation (JSON) file, which is a standard tool for capturing structure information.

C. Compared Methods

We compare our model with a mass of baseline and state-of-the-art approaches: VGG-16 Network [70], LRCN [71], Multi-target Tracking [26], HDTM [1], HiRF [5], Contextual Model [3], CERN [33], Cardinality Kernel [15], Deep Structure Model [11], Recurrent Modeling [34],SBGAR [36], SRNN [37], SSU [35], $E^2(\infty)$ [7] and V1(∞) [74][1].

In particular, as shown in Table I, 'VGG-16-Image' and 'LRCN-Image' extract the holistic image features of every single frame for recognition. 'VGG-16-Person' and 'LRCN-Person' distinguish group activities using features pooled over all cropped-size person-level features. 'HDTM' and 'CERN' methods conduct experiments utilizing the grouping strategy on the Volleyball Dataset, which splits all players into one or two groups, and 'SRNN' also divides persons into two groups. 'Recurrent Modeling' and 'SBGAR' utilize two kinds

[1]In the experiments, the parameter setting of above-mentioned methods are adopted from the corresponding papers.

TABLE I
PERFORMANCE COMPARISON OF OUR METHOD AND THE STATE-OF-THE-ART APPROACHES. 'SEMANTIC' INDICATES WHETHER THE METHOD CAN EXTRACT AND OUTPUT SEMANTIC INFORMATION. 'ACCURACY-C' SHOWS THE GROUP ACTIVITY RECOGNITION ACCURACIES ON COLLECTIVE ACTIVITY. 'ACCURACY-N' SHOWS THE GROUP ACTIVITY RECOGNITION ACCURACIES ON NEW COLLECTIVE ACTIVITY. 'ACCURACY-V-1' AND 'ACCURACY-V-2' RESPECTIVELY DEPICT THE GROUP ACTIVITY AND PERSONAL ACTION RECOGNITION ACCURACIES ON VOLLEYBALL. 'PROPOSAL' AND 'GT' INDICATE THAT WE USE BOUNDING BOXES OBTAINED BY PROPOSAL AND GROUND-TRUTH BOUNDING BOXES PROVIDED BY [1], RESPECTIVELY. THE BEST PERFORMANCE IS HIGHLIGHTED IN RED AND THE SECOND BEST IN BLUE.

Methods	Semantic	Accuracy-C	Accuracy-N	Accuracy-V-1	Accuracy-V-2
VGG-16-Image [70]	×	68.3	71.9	71.7	-
VGG-16-Person [70]	×	71.2	75.3	73.5	-
LRCN-Image [71]	×	64.2	69.3	63.1	-
LRCN-Person [71]	×	64.0	69.1	67.6	-
HDTM (1 group) [1]	×	81.5	-	70.3	75.9
HDTM (2 groups) [1]	×	-	-	81.9	-
Multi-target Tracking [26]	×	79.6	83.0	-	-
HiRF [5]	×	83.1	87.3	-	-
Contextual Model [3]	×	79.1	-	-	-
Deep Structure Model [11]	×	81.2	89.5	-	-
Cardinality kernel [15]	×	83.4	-	-	-
CERN-1 (1 group) [33]	×	84.8	-	34.4	69.0
CERN-2 (1 group) [33]	×	87.2	-	73.5	-
CERN-2 (2 groups) [33]	×	-	-	83.3	-
Recurrent Modeling (RGB&Optical Flow) [34]	×	89.4	85.2	-	-
SBGAR (RGB&Optical Flow) [36]	√	86.4	-	67.7	-
SRNN (2 groups) [37]	×	-	-	83.5	76.6
SSU-temporal (MRF) [35]	×	-	-	87.1	-
SSU-temporal (GT) [35]	×	-	-	89.9	82.4
Ours w/o attention (Proposal)	√	85.6	87.5	85.7	79.6
Ours w/ attention (Proposal)	√	87.9	89.2	87.6	-
Ours w/o attention (GT)	√	87.7	89.6	87.9	81.9
Ours w/ attention (GT)	√	89.1	90.2	89.3	82.3

of features by the spatial ConvNets (AlexNet [75]) for original images and the motion ConvNets (GoolgeNet [76]) for flow images. 'SSU-temporal' models employ two types of detection methods on the Volleyball Dataset, *i.e.* the ground truth bounding boxes (GT) and Markov Random Fields (MRF) based detection. **Note that** 'LRCN', 'HDTM', 'SRNN' and 'Deep Structure Model' employ the AlexNet [75] as the backbone, and 'SSU' and 'SBGAR' adopt the Inception-V3 [77] framework, while 'CERN' and our model select the VGG-16 model. Besides, 'Multi-target Tracking', 'HiRF', 'Contextual Model', 'Cardinality Kernel', '$E^2(\infty)$' and 'V1(∞)' extract hand-crafted visual features rather than deep learning based approaches. Only $E^2(\infty)$ [7] and V1(∞) [74] have conducted experiments on UCLA Courtyard dataset. $E^2(\infty)$ utilized a cost-sensitive explore-exploit strategy to optimize with the hierarchical AND-OR graph, and V1(∞) employed a Monte Carlo Tree Search as inference method with an expressive AND-OR graph.

D. Results and Analysis

Results on the Collective Activity Dataset. The experimental results are shown in the 'Accuracy-C' column of Table I. As can be seen, our model with the attention model obtains the second-best performance among the compared state-of-the-art methods, no matter using the proposal-based or ground-truth bounding boxes. As an example, our model gains $\approx$15% higher in accuracy than VGG/LRCN that are image-level and person-level classification methods, mostly attributed to our RNN-based framework with the iteratively message passing and factor sharing scheme. Additionally, the improved result demonstrates that the spatio-temporal semantic graph is beneficial for improving the recognition accuracy. Note that 'Recurrent Modeling' obtains the best performance by adopting two varieties of features, *i.e.* RGB and optical flow, while our stagNet only employ RGB visual features. Meanwhile, our method and 'SBGAR' are the only two that incorporates semantics into the model. However, our method is only one to output graph structure based information. while 'SBGAR' utilize sentence based representations. The cardinality kernel approach [15] obtains the best performance among non-deep learning methods (with hand-crafted features) by counting the numbers of individual actions in a frame directly. Additionally, Fig. 9(a) illustrates the confusion matrix based on our stagNet with spatio-temporal attention mechanism. We can find that nearly 100% recognition accuracies in terms of 'queuing' and 'talking', demonstrating the effectiveness of our model. However, quite a few failure cases still exist because some action classes share high similarities, *e.g.* 'walking' and 'crossing'. If we have more training data of different action categories, the classification accuracy will be improved.

Results on New Collective Activity Dataset. We also draw experiments with New Collective Activity Dataset. The 'Accuracy-N' column of Table I illustrates that the group activity classification results. As can be seen, our model outperforms the baseline and the state-of-the-art approaches, regardless with attention or without attention mechanism. Especially, our model gains a 15% improvement in group activity recognition accuracy over deep learning baseline methods (*i.e.* VGG-16 and LRCN), mostly because of the additional modeling of inter-person relationships. In addition, we draw the confusion matrix based on our model in Fig. 9(b). Nearly 100% recognition accuracies can be obtained in terms of

'talking' and 'queuing', and the accuracies of other categories are more than 70%. However, the accuracy of 'gathering' and 'dismissal' is not good enough, since 'gathering' and 'talking' have the similar visual appearance in a crowd scene, and 'dismissal' class is easy being confused by 'talking' and 'walking' when these classes activities occurred in the same time.

Results on UCLA Courtyard. We have conducted extra experiments on UCLA Courtyard [7]. As shown in Table II and Table III, our method consistently outperforms existing methods in terms of group activity and individual action recognition. For each activity category, our stagNet gains about more than 15% and 3% than $E^2(\infty)$ and V1(∞), respectively. Moreover, our method achieves 86.3% in terms of average activity recognition accuracy that is the best performance in UCLA Courtyard dataset. As for individual action recognition, our stagNet achieves the best performance across all action categories than another two state-of-the-art methods. The reason why our proposed method outperforms the other approaches is that our stagNet can learn far better spatio-temporal representation through semantic graph architecture, and attention mechanism presented in stagNet is beneficial to the personal action recognition.

TABLE II
THE ACCURACY OF GROUP ACTIVITY RECOGNITION WITH OUR PROPOSED MODEL AND THE STATE-OF-THE-ART METHODS ON UCLA COURTYARD. BEST RESULTS ARE IN BOLD.

Group Activity	$E^2(\infty)$ [7]	V1(∞) [74]	stagNet
Standing-in-line	68.0	80.4	**83.1**
Guided-tour	70.2	83.5	**86.5**
Discussing	75.1	81.5	**83.6**
Sitting	71.4	87.2	**89.1**
Walking	78.6	88.6	**89.5**
Waiting	72.6	80.1	**83.3**
Average	N/A	83.7	**86.9**

TABLE III
AVERAGE PRECISION AND FALSE POSITIVE RATES (IN BRACKETS) OF INDIVIDUAL ACTION RECOGNITION ON UCLA COURTYARD DATASET OF OUR PROPOSED MODEL AND THE STATE-OF-THE-ART METHODS. BEST RESULTS ARE IN BOLD.

Action	$E^2(\infty)$ [7]	V1(∞) [74]	stagNet
Walk	69.1(18.7)	80.0(17.1)	**82.1(16.5)**
Wait	67.7(20.2)	80.0(18.8)	**82.9(17.0)**
Talk	69.6(17.9)	76.8(16.6)	**79.2(15.5)**
Drive	70.2(9.7)	82.1(8.1)	**85.5(7.3)**
Surf	71.3(17.1)	79.8(15.4)	**81.6(13.3)**
Scoot	68.4(16.3)	81.8(14.1)	**82.5(12.5)**
Bike	61.4(12.3)	76.9(12.2)	**78.3(11.9)**
Read	67.3(12.1)	79.6(10.1)	**81.7(9.2)**
Eat	71.3(7.7)	82.3(6.5)	**83.9(5.6)**
Sit	64.2(9.0)	75.5(8.1)	**77.9(7.5)**

Results on the Volleyball Dataset. The recognition results of proposed stagNet and the state-of-the-art are displayed in the 'Accuracy-V-1/2' columns of Table I. We can see that the group activity and personal action recognition results of our model is superior to most of the state-of-the-art approaches, and additionally competitive to the best 'SSU' extremely. It ought to be noted that 'SSU' achieves the bounding boxes by a much more sophisticated multi-scale technique and employs the more advanced Inception-V3 model as the backbone. While our stagNet adopt the basic VGG-16 and the 'ground-truth' bounding boxes provided by [1] without other advanced object detection strategy. Therefore, it can be expected that the performance of our method could be further improved by employing more advanced object detection models and backbone networks. Additionally, our model outperforms other RNN-based approaches (*e.g.* HDTM/CERN/SRNN) by about $5 \sim 8\%$ w.r.t. group activity recognition, because semantic graph with structural-RNN in our model can extract and model better spatio-temporal relationships. And integrating the spatio-temporal attention model can further improve performance, suggesting that different personal action plays the various role and the most crucial persons' visual features are significant for recognizing the whole group activity. As for individual action recognition, our proposed stagNet accomplish the second best performance (just lower 1% than SSU), indicating the effectiveness of the body-region attention mechanism and global-part pooling strategy in our method. In addition, it is also worth noting that all the other state-of-the-art are unable to capture the semantic structural information for describing the scene. In contrast, our proposed approach can describe the semantic contents of the scene via outputting a semantic graph. Fig. 6 illustrates the recognition results and the corresponding semantic graphs visually. Moreover, Fig. 9(c) showed the confusion matrix based on our stagNet. For the majority of group activities, we can reach promising recognition accuracies ($\geq$87%).

Parameter Sensitivity Analysis. Parameters have a great influence on experiments performance. Therefore, we assess the impact of two parameters when training our model, *i.e.* Training Epochs and Sliding Window Size. We define each epoch as the process input the entire training set for training model each time. And the sliding window size is defined as the number of video frames used to output a recognition result. As shown in Fig. 7(a), we report the relationship between the accuracy of group activity recognition and the number of training epochs. We can observe that the more epochs times result in higher accuracy, and when epochs exceed 600 the accuracy tends to stable in all dataset. Fig. 7(b) shows the group accuracy on all dataset when select different sliding window size (*i.e.* 5/10/15/20 frames). It is apparent that employing a sliding window of 10 frames, we can get the best accuracy.

In addition, there are totally five parameters in the objective function of our model: K (number of players), T (number of frames), λ_1, λ_2 and λ_3 (see Eq. (10)). The value of K depends on the specific dataset (*e.g.* K=12 w.r.t. Volleyball). Fig. 8 shows the sensitivity of λ_1, λ_2 and λ_3 on all the four benchmark datasets. Through the experimental results, we set $\{\lambda_1,\lambda_2,\lambda_3\}$ as $\{0.001,0.0001,0.0001\}$ for Collective Activity/New Collective Activity, and $\{0.01,0.001,0.00001\}$ for UCLA Courtyard/Volleyball for the best performance in practice.

Qualitative Results Analysis. Fig. 13, Fig. 14, and Fig. 15 visually depict the individual and group activity recognition results, and the attention paid to different subjects in the scene using heat maps in terms of proposal and ground-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2894161, IEEE Transactions on Circuits and Systems for Video Technology

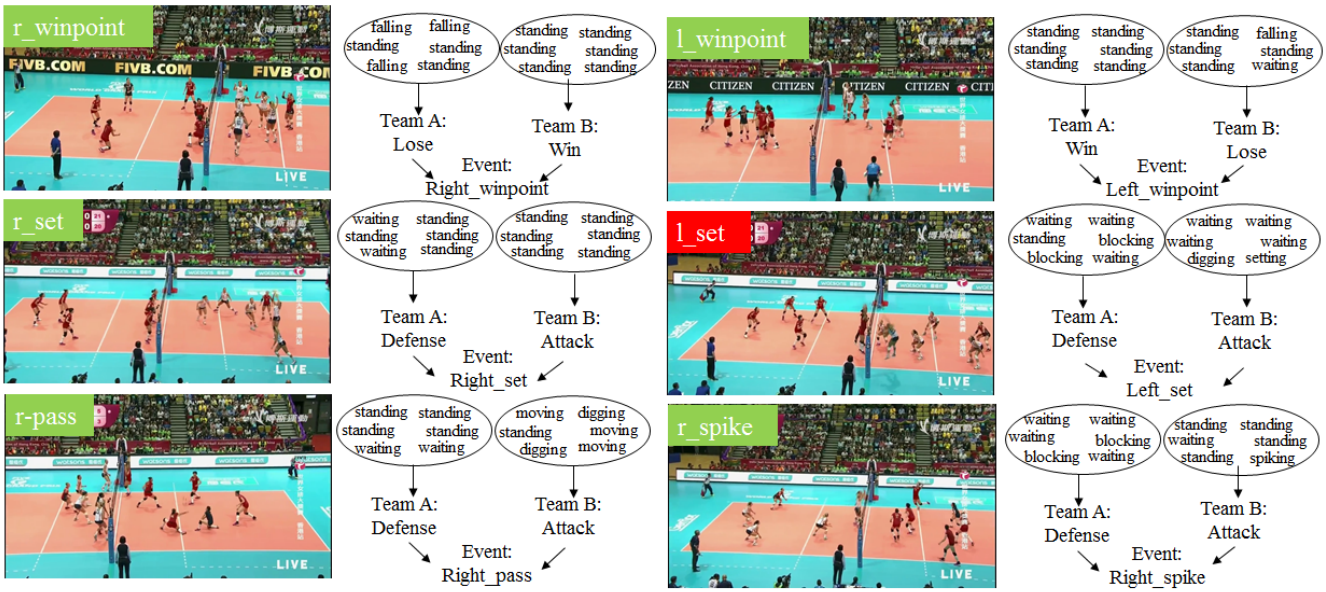
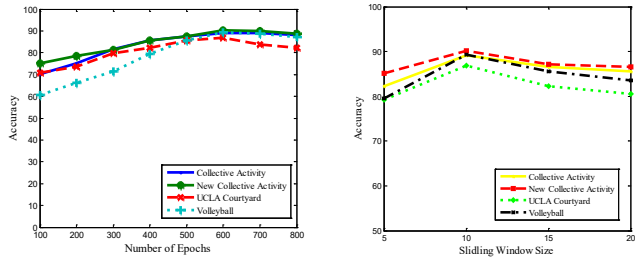


Fig. 6. Visualization of group activity recognition results. Green texts denote successful results, and red ones indicate failure cases. The corresponding semantic graphs obtained by our method are shown to explain the context of the whole scene. Such a graph generally consists of person actions, team activities and group activity (*i.e.* event).



(a) Numer of Epochs (b) Sliding Window Size

Fig. 7. Recognition accuracy w.r.t. number of epochs and sliding window size on the four datasets.

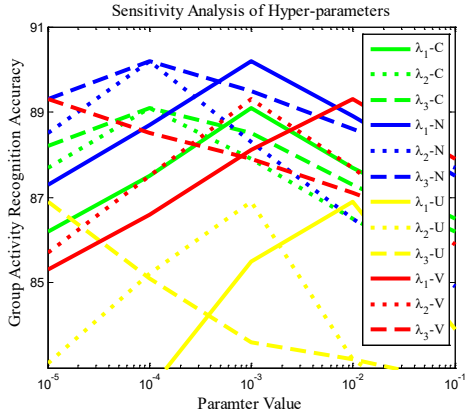


Fig. 8. Parameter sensitivity analysis on Collective Activity (C), New Collective Activity (N), UCLA Courtyard (U) and Volleyball (V).

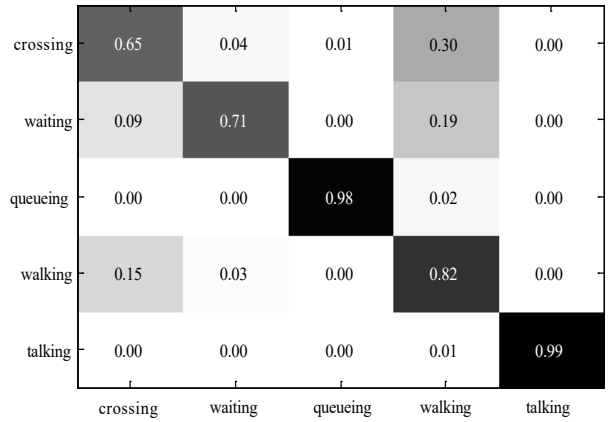
truth (GT) bounding boxes. We draw the heat maps based on the attention weights for objects/persons in each frame. We can observe from the figures that some individual actions play significant roles in distinguishing the corresponding high-level group activity, such as individual 'standing' w.r.t. group 'queueing', individual 'running' w.r.t. group 'chasing', and individual 'spiking/blocking' w.r.t. group 'left-spike'. This further indicates that the spatio-temporal attention model is necessary and beneficial for recognizing the overall group activity. In addition, we show some failure cases in the bottom right corner of each figure, which is probably due to that some action classes share high similarities with each other. For instance, 'walking' in Fig. 13 is mistaken as 'crossing' probably because they both involve the individual actions 'front' and 'back', and similar street background. In Fig. 14, 'dismissal' is misled by 'walking together' due to that the individual 'walking' commonly exists in both categories, and there is no more useful label information in the dataset. In Fig. 15, 'L-pass' is mistaken as 'L-spike', since 'pass' looks very similar to 'spike' with hands up in the air. This also indicates that the position of 'ball' is important for activity recognition. In order to distinguish such ambiguous activities, more training data or annotations will be required.

Additionally, Fig. 10 and Fig. 11 illustrate the visualization of temporal attention for videos and body region attention for the individual player. As can we see from Fig. 10, given a video of 'Left spike' on Volleyball, the 5-th frame is the most significant than the others as a result of a player is spiking and two players are trying to block at the same time. These movements of key players determine the final group activity recognition result considerably. Besides, Fig. 11 depicts six examples of body region attention heat map. We can notice that the regions of arms and legs invariably play a paramount role to affect the individual action. As an example, the arm in upper right and legs in the bottom right in '(c)spiking' are given the higher attention weights than anther body regions. It clearly demonstrates that our body-region attention mechanism is reasonable and necessary.

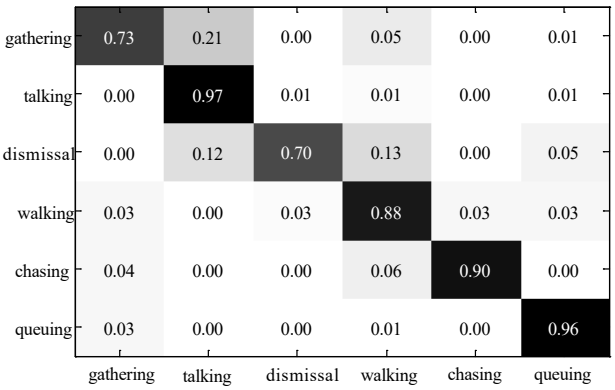
E. Ablation Study

In this subsection, we perform several ablation studies to better examine the effect of our proposed stagNet.

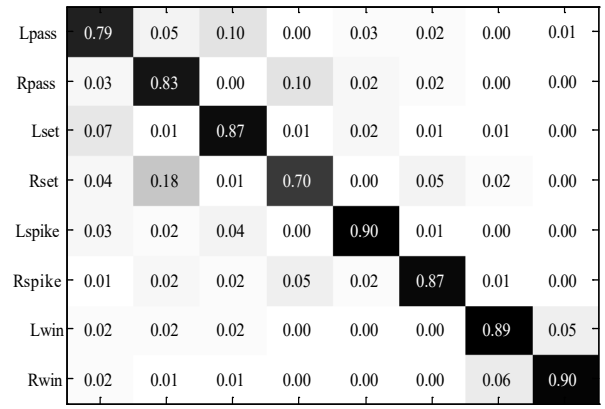
Body-Region Attention and Global-Part Pooling. To explore the effectiveness of our proposed "body-region attention" and "Global-Part Pooling" mechanism, we perform



(a) Collective Activity



(b) New Collecitve Activity



(c) Volleyball

Fig. 9. Confusion matrices for three group activity datasets in our experiments.

the ablative experiments and analysis. For comparison, we introduce four variants in our experiments for individual action recognition, *i.e.* "Max-Pooling", "Average-Pooling", "Only Global", and "Only Attention", which refer to utilizing maximized body region feature, averaged all the body regions' features, only global person bounding box feature without body-region attention, only body-region attention without global-part pooling, respectively. Table IV illustrates the quantitative results on the Volleyball Dataset [1]. Obviously, we can find that our full model, combining body-region attention with global-part pooling together, outperforms other variant methods, and achieves the best performance. In contrast, "Max-Pooling" and "Average-Pooling" achieve poor results, suggesting global structural information of individual player is significant for action recognition. Clearly, we can draw the conclusion that our full model is effective attributed to taking each body region feature attentively and global player body information into consideration simultaneously. Furthermore, we examine the effect of number of body regions in our model in terms of accuracy and computation time for individual action recognition. As shown in Fig. 12, the accuracy is improved and computation time is augmented along with the increasing of body regions numbers. However, the growth rate of performance is considerably slower when the number of body regions we selected exceeds six, while the increasing rate of time consuming is still high. This illustrates much more body splits are not economical. Therefore, we set the number of body regions as six in our proposed model to obtain the trade-off balance. Additionally, it is promising to incorporate more pose information into our model as our future work for sports video captioning [78].

TABLE IV
EXPERIMENT RESULTS FOR INDIVIDUAL ACTION RECOGNITION ON VOLLEYBALL [1] DATASET. **Max**, **Ave**, **Glo**, **Att** AND **Ours-Full** DENOTE MAX-POOLING, AVERAGE-POOLING, ONLY GLOBAL, ONLY ATTENTION AND OUR FULL MODEL, RESPECTIVELY. BEST RESULTS ARE IN BOLD.

Performance	Max	Ave	Glo	Att	Ours-Full
mAP(%)	73.6	77.6	81.9	82.0	**82.3**

Potential Efficiency. Most of existing methods invariably neglect the balance between computational cost and recognition accuracy. Therefore, it is necessary to reduce the computational cost for untrimmed video classification while retaining reasonable accuracy. Following the idea in [79], we incorporate the method in the paper to examine the potential efficiency of our model for group acitivity and individual action recognition on Volleyball Dataset [1]. The method in the paper proposed an end-to-end deep reinforcement approach which introduces an agent to classify videos only by watching a small portion of frames, based on "fast forward" and "adaptive stop" mechanism. As can be illustrated in Fig. 7(b), the accuracy can improve along with increasing watched frames, while the accuracy can not improve much when the number of the watched frames exceeds 10. Furthermore, Table V lists the results of our original stagNet and stagNet incorporated with RL-LSTM/RL-GRU [79]. We evaluate our original stagNet in two cases: randomly sampling 10 frames (R10), and uniformly sampling 10 frames (U10). Meanwhile, the setting of RL-LSTM/RL-GRU is followed in [79]. All eight group activity categories are utilized in our experiments. As shown in Table V, our stagNet with RL-GRU achieves the best performance, which only watches 8.66 frames on average but obtains 89.9% in terms of mAP. It demonstrates that incorporating the method [79] into our stagNet is able to improve the recnognition performance and efficiency. Additionally, the nodeRNNs and edgeRNNs in our proposed stagNet need to be applied for every frame. However, we introduce the temporal attention to assign different important scores to every frame, so that we can select the critical frame from a video to determine the group activity correctly, which also has potential efficiency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2019.2894161, IEEE Transactions on Circuits and Systems for Video Technology

Fig. 10. Visualization of temporal attention for 'Left-spike' group activity on Volleyball Dataset. The attention weights vary from large to small as long as the colors changing from red to blue.

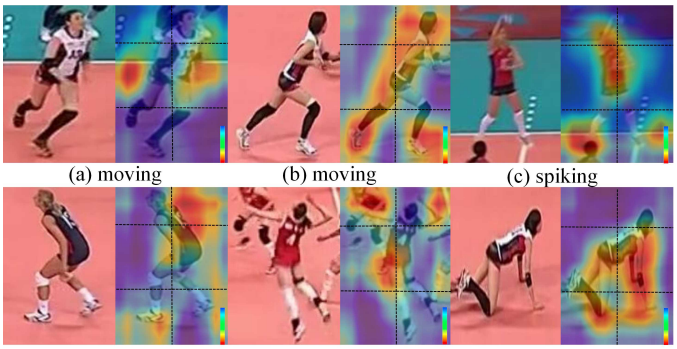


Fig. 11. Visualization of body-region attention for individual player on Volleyball. The attention weights change from large to small along with the colors changing from red to blue.

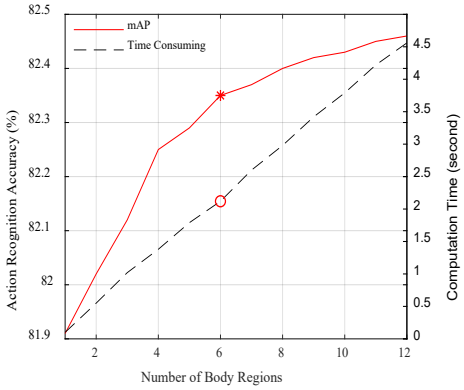


Fig. 12. Parameter sensitivity analysis of the body region number for individual action recognition on Volleyball [1].

for video-based recognition tasks.

TABLE V
EXPERIMENT RESULTS (MAP) OF OUR PROPOSED ORIGINAL STAGNET AND OUR STAGNET WITH RL-LSTM/RL-GRU [79] FOR GROUP ACTIVITY RECOGNITION ON VOLLEYBALL DATASET [1]. **R10** AND **U10** DENOTE RANDOMLY SAMPLING 10 FRAMES AND UNIFORMLY SAMPLING 10 FRAMES, RESPECTIVELY. BEST RESULTS ARE IN BOLD.

Category	stagNet		RL-LSTM [79]		RL-GRU [79]	
	R10	U10	mAP(%)	#frames	mAP(%)	#frames
L-pass	79.1	79.2	79.3	6.97	**80.5**	8.37
R-pass	81.5	83.1	84.5	6.52	**85.0**	8.76
L-set	86.9	87.2	87.3	6.15	**87.6**	7.79
R-set	70.5	70.3	71.2	8.05	**71.9**	8.28
L-spike	90.3	90.2	90.5	8.23	**90.8**	7.52
R-spike	87.2	87.5	87.9	7.06	**88.6**	9.27
L-win	89.0	89.1	88.7	6.68	**90.3**	8.69
R-win	89.9	90.3	89.6	6.93	**91.7**	8.15
mean	89.1	89.3	89.2	7.25	**89.9**	8.66

Different Training Data and Unsupervised Training. Because the current datasets for group activity recognition are relatively small, it is meaningful to exploit the effect of pre-training model with larger training data (*e.g.* Kinetics [80] and ActivityNet [81]) or unsupervised manner [82] for our proposed stagNet. The reason why we choose pre-trained VGG16 on ImageNet in our experiments is to make fair comparisons with other state-of-the-art methods. Because our task mainly focuses on person activity and action recognition, we firstly conduct ablative experiment with pre-trained model (ConvNet+LSTM) on Kinetics datasets that is a large-scale human action video dataset. Moreover, in order to overcome the lack of sufficient labeled data, we perform the experiment by adopting the pre-trained model in an unsupervised way based on Zhu et al. [82]. The method of the paper is to learn multirate representations for videos via context reconstruction in an unsupervised training. Table VI shows the results of our proposed stagNet pre-trained on Kinetics compared with that pre-trained on ImageNet, and in an unsupervised training strategy. From the table, we can find that our stagNet pre-trained on Kinetics can obtain the better performance than that pre-trained on ImageNet, demonstrating that adopting a pre-trained model on such a large-scale action video dataset for activity/action recognition is necessary and beneficial for our task. In addition, our proposed stagNet with unsupervised training strategy in [82] obtains highly competitive performances, suggesting the unsupervised learning is promising to solve the difficulty of insufficient annotated data.

TABLE VI
THE ACCURACY OF GROUP ACTIVITY AND INDIVIDUAL ACTION(IN BRACKETS) RECOGNITION ON COLLECTIVE ACTIVITY [23] (COL)/NEW COLLECTIVE ACTIVITY [26] (NEW)/VOLLEYBALL [1] (VOL) DATASETS WITH PRE-TRAINED MODEL IN DIFFERENT DATASETS AND UNSUPERVISED TRAINING STRATEGY. BEST RESULTS ARE IN BOLD.

Methods	Col [23]	New [26]	Vol [1]
stagNet w/ ImageNet [83]	89.1	90.2	89.3(82.3)
stagNet w/ Kinetics [80]	**90.2**	**92.5**	**90.5(83.5)**
stagNet w/ Unsupervised [82]	88.3	88.9	87.6(80.0)

F. Discussion

Above all, our framework based on the semantic graph and spatial-temporal attention mechanism for group activity and individual action recognition is feasible in modeling the inter-person relationship by aggregating time and space features. Firstly, it is demonstrated that semantic graph based RNN architecture has advantages in temporal feature aggregation and context information extraction. Modeling *group-person interaction* and *person-person* interaction will improve understanding group activity and individual action in sports match and surveillance videos. And it is promising to utilize more complicate dynamic scene and large-scale variety data stream. Besides, the structural semantic output is beneficial for lots of other tasks like dense video captioning [84], sports video captioning [78] and visual question answering [85] as it provides mid-level relationships for fine-grained recognition.

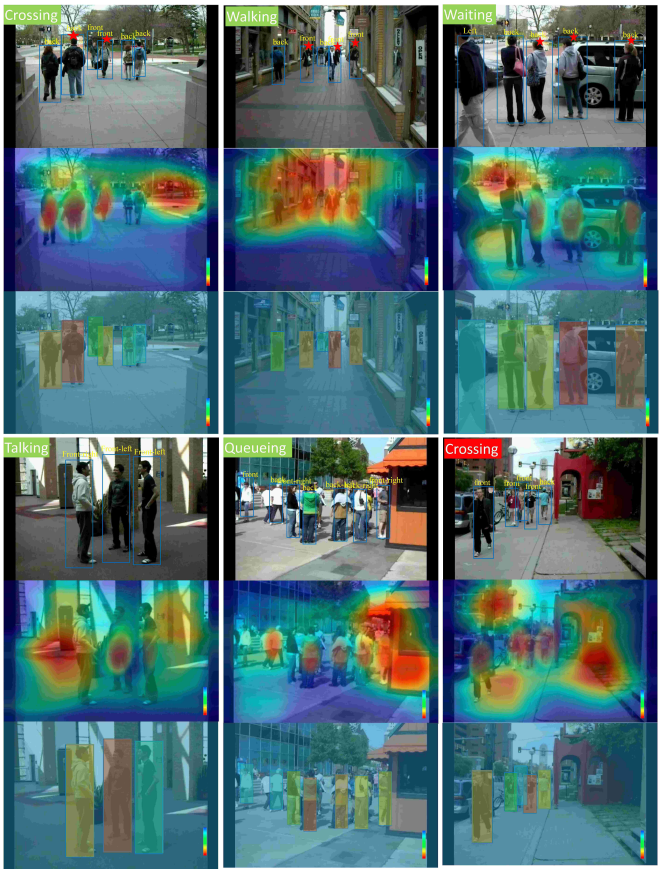


Fig. 13. Visualization results on the Collective Activity dataset. Group activity and individual action recognition results are shown in the first/fourth row, and attention heat maps based on proposal and ground-truth (GT) bounding boxes are shown in the second/fifth row and the third/sixth row, respectively. Green texts indicate successful results, and red ones are failure cases. The important persons in the scene are denoted with red stars. The attention weights change from large to small along with the colors changing from red to blue.

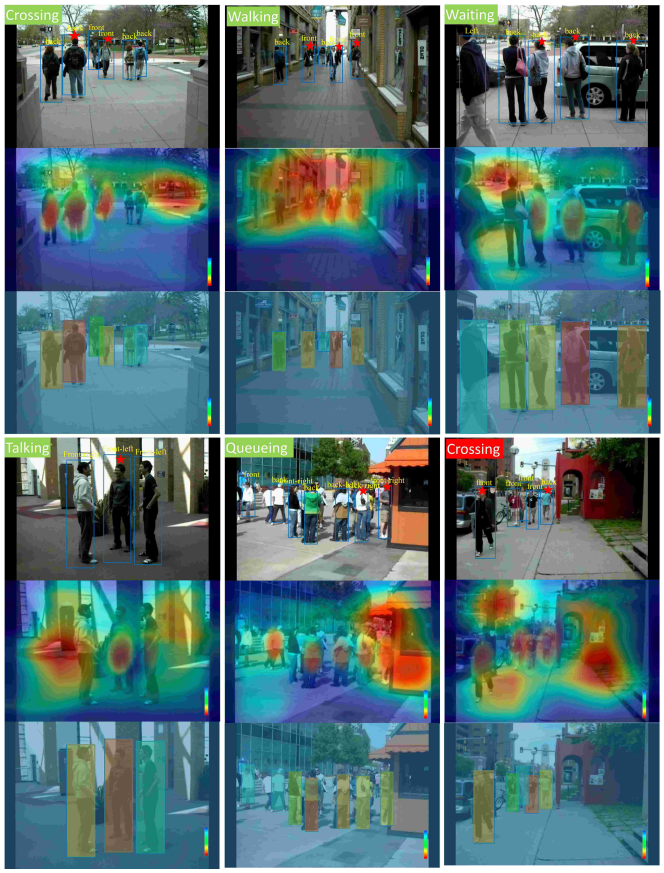


Fig. 14. Visualization results on the New Collective Activity dataset. The other settings are the same as in Fig. 13

Moreover, it is necessary and promising to apply actively sampling informative data for improved action recognition performance. The existing public datasets are limited and the number of labeled examples is usually small. However, massive data can be achieved on the Internet, such as sports games and surveillance videos. To relieve the tedious work of manual annotation and exploit the uncertainty across multiple classes, multi-class active learning [86] is able to select the most informative data from a candidate set for labeling, and decides what data are more helpful and then asks humans to label them for training. If we can adopt such multi-class active learning method into our issue, we believe that our model can obtain better performance and capture more meaningful representations from large-scale data. We leave this for our future work, which is expected to further enhance the overall performance.

Last but not least, most related works and our model specialize in group activity recognition rather than object detection, hence we utilize mainstream object detection models (*e.g.* Faster RCNN, YOLO). Obviously, it will be helpful to handle the situations (*e.g.* occlusion and people leaving the scene) if we employ more accurate objection detection models.

V. Conclusion

In this paper, we propose a novel Recurrent Neural Network with semantic graph and spatio-temporal attention mechanism, named as *stagNet*, for group activity and individual action recognition. Our framework could capture spatio-temporal representation and inter-object relationships in a dynamic scene with a semantic graph explicitly. Through the inference procedure by virtue of message passing between nodeRNNs and edgeRNNs, our model is capable of predicting the label of the whole scene, each individual action, and inter-person interaction at the same time. By incorporating the spatio-temporal attention mechanism into our proposed framework further, important persons, body regions or frames in the video can be concentrated on, resulting in better recognition performance. Extensive experimental results across four widely-adopted public benchmarks demonstrate that our approach acquires competitive performance to the state-of-the-art, whilst outputting the detailed semantic description of the scene with a structural graph. Future work will investigate new tools and techniques about unsupervised learning so that we can utilize and learn better features from unlabeled activity categories. Besides, we will improve our framework by integrating reinforcement learning strategy.

References

[1] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in *CVPR*. IEEE, 2016, Conference Proceedings, pp. 1971–1980.

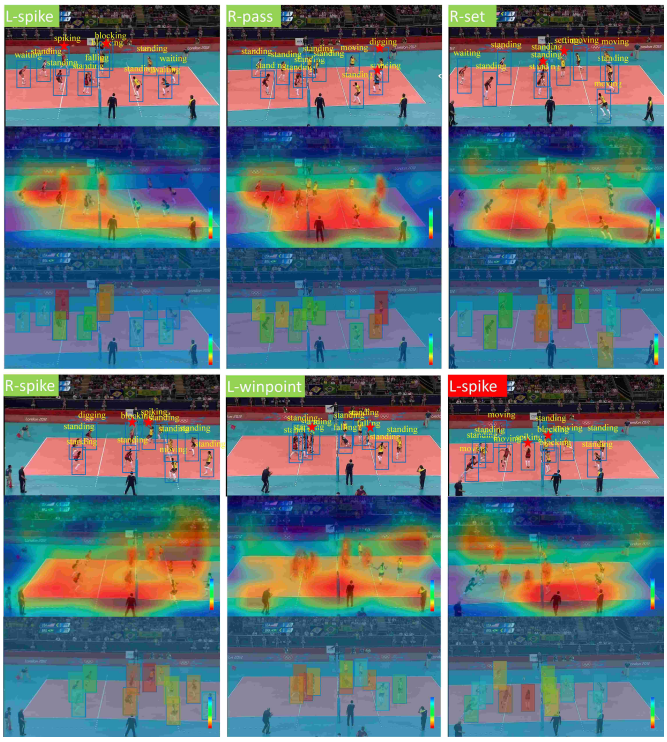


Fig. 15. Visualization results on the Volleyball dataset. The other settings are the same as in Fig. 13

[2] W.-L. Lu, J.-A. Ting, J. J. Little, and K. P. Murphy, "Learning to track and identify players from broadcast sports videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1704–1716, 2013.

[3] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549–62, 2012.

[4] Z. Wang, Q. Shi, C. Shen, and V. D. H. Anton, "Bilinear programming for human activity recognition with unknown mrf graphs," in *CVPR*. IEEE, 2013, Conference Proceedings, pp. 1690–1697.

[5] M. R. Amer, P. Lei, and S. Todorovic, "Hirf: Hierarchical random field for collective activity recognition in videos," in *ECCV*. Springer International Publishing, 2014, Conference Proceedings.

[6] T. Shu, D. Xie, B. Rothrock, and S. Todorovic, "Joint inference of groups, events and human roles in aerial videos," in *CVPR*. IEEE, 2015, Conference Proceedings, pp. 4576–4584.

[7] M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S. C. Zhu, "Cost-sensitive top-down/bottom-up inference for multiscale activity recognition," in *ECCV*. Springer International Publishing, 2012, Conference Proceedings, pp. 187–200.

[8] J. Liu, P. Carr, R. T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *CVPR*. IEEE, 2013, Conference Proceedings.

[9] J. Qin, L. Liu, L. Shao, F. Shen, B. Ni, J. Chen, and Y. Wang, "Zero-shot action recognition with error-correcting output codes," in *CVPR*. IEEE, 2017, Conference Proceedings.

[10] J. Qin, L. Liu, L. Shao, B. Ni, C. Chen, F. Shen, and Y. Wang, "Binary coding for partial action analysis with limited observation ratios," in *CVPR*. IEEE, 2017, Conference Proceedings.

[11] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *CVPR*. IEEE, 2016, Conference Proceedings.

[12] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *CVPR*. IEEE, 2016, Conference Proceedings, pp. 5308–5317.

[13] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*. Springer, 2006, Conference Proceedings, pp. 428–441.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1. IEEE, 2005, Conference Proceedings, pp. 886–893.

[15] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, "Visual recognition by counting instances: A multi-instance cardinality potential kernel," in *CVPR*. IEEE, 2015, Conference Proceedings.

[16] Y. Bengio, Y. LeCun, and D. Henderson, "Globally trained handwritten word recognizer using spatial representation, convolutional neural networks, and hidden markov models," in *NeurIPS*, 1994, Conference Proceedings, pp. 937–944.

[17] P. Krhenbhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NeurIPS*, 2011, Conference Proceedings.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, Conference Proceedings, pp. 1097–1105.

[19] R. S, H. K, G. R, and S. J, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, p. 1137, 2017.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE, 2016, Conference Proceedings, pp. 770–778.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv*, 2014.

[23] W. Choi, K. Shahid, and S. Savarese, "What are they doing? : Collective activity classification using spatio-temporal relationship among people," in *ICCV Workshops*. IEEE, 2009, Conference Proceedings, pp. 1282–1289.

[24] W. Choi and K. Shahid, "Learning context for collective activity recognition," in *CVPR*. IEEE, 2011, Conference Proceedings, pp. 3273–3280.

[25] M. Qi, J. Qin, A. Li, Y. Wang, J. Luo, and L. Van Gool, "stagnet: An attentive semantic rnn for group activity recognition," in *ECCV*. Springer, 2018, Conference Proceedings.

[26] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *ECCV*. Springer International Publishing, 2012, Conference Proceedings, pp. 215–230.

[27] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *NeurIPS*, 2016, Conference Proceedings.

[28] S. Khamis, V. I. Morariu, and L. S. Davis, "Combining per-frame and per-track cues for multi-person action recognition," in *ECCV*. Springer, 2012, Conference Proceedings.

[29] S. Kwak, B. Han, and J. H. Han, "Multi-agent event detection: Localization and role assignment," in *CVPR*. IEEE, 2013, Conference Proceedings, pp. 2682–2689.

[30] M. S. Ryoo and J. K. Aggarwal, "Stochastic representation and recognition of high-level group activities: Describing structural uncertainties in human activities," *International Journal of Computer Vision*, vol. 93, no. 2, pp. 183–200, 2011.

[31] G. Mori, "Social roles in hierarchical models for human activity recognition," in *CVPR*. IEEE, 2012, Conference Proceedings, pp. 1354–1361.

[32] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.

[33] T. Shu, S. Todorovic, and S. C. Zhu, "Cern: Confidence-energy recurrent network for group activity recognition," in *CVPR*. IEEE, 2017, Conference Proceedings.

[34] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *CVPR*. IEEE, 2017, Conference Proceedings.

[35] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *CVPR*. IEEE, 2017, Conference Proceedings.

[36] X. Li and M. C. Chuah, "Sbgar: Semantics based group activity recognition," in *ICCV*. IEEE, 2017, Conference Proceedings.

[37] S. Biswas and J. Gall, "Structural recurrent neural network (srnn) for group activity analysis," *arXiv preprint arXiv:1802.02091*, 2018.

[38] M. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in *ECCV*, 2018.

[39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*. IEEE, 2016, Conference Proceedings, pp. 779–788.

[40] S. E, L. J, and D. T, "Fully convolutional networks for semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 4, p. 640, 2017.

[41] X. Li and M. C. Chuah, "Rehar: Robust and efficient human activity recognition," *arXiv preprint arXiv:1802.09745*, 2018.

[42] H. Fan, X. Chang, D. Cheng, Y. Yang, D. Xu, and A. G. Hauptmann, "Complex event detection by identifying reliable shots from untrimmed videos," in *ICCV*. IEEE, 2017, Conference Proceedings.

[43] M. Qi, Y. Wang, and A. Li, "Online cross-modal scene retrieval by binary representation and semantic graph," in *MM*. ACM, 2017, Conference Proceedings.

[44] A. Krogh, B. Larsson, G. Von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1," *Journal of molecular biology*, vol. 305, no. 3, pp. 567–580, 2001.

[45] L. C. Chen, A. G. Schwing, A. L. Yuille, and R. Urtasun, "Learning deep structured models," *ICLR*, pp. 1785–1794, 2014.

[46] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang, "Semantic image segmentation via deep parsing network," in *ICCV*. IEEE, 2015, Conference Proceedings, pp. 1377–1385.

[47] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *CVPR*. IEEE, 2014, Conference Proceedings, pp. 1637–1644.

[48] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3d human pose estimation," in *ICCV*. IEEE, 2015, Conference Proceedings, pp. 2848–2856.

[49] J. Tompson, A. Jain, Y. Lecun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *NeurIPS*, 2014, Conference Proceedings.

[50] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arxiv*, no. 4, pp. 357–361, 2014.

[51] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *ICCV*. IEEE, 2015, Conference Proceedings.

[52] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee, "Improving object detection with deep convolutional networks via bayesian optimization and structured prediction," in *CVPR*. IEEE, 2015, Conference Proceedings, pp. 249–258.

[53] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *NeurIPS*, vol. 4, pp. 2951–2959, 2012.

[54] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *NeurIPS*, 2016, Conference Proceedings, pp. 3844–3852.

[55] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs," in *ICML*, 2016, Conference Proceedings, pp. 2014–2023.

[56] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[57] N. Shapovalova, M. Raptis, L. Sigal, and G. Mori, "Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization," in *NeurIPS*, 2013, Conference Proceedings.

[58] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *NeurIPS*, 2014, Conference Proceedings.

[59] C. Cao, X. Liu, Y. Yang, and Y. Yu, "Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks," in *ICCV*. IEEE, 2015, Conference Proceedings.

[60] J. C. Caicedo and S. Lazebnik, "Active object localization with deep reinforcement learning," in *ICCV*. IEEE, 2015, Conference Proceedings.

[61] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015, Conference Proceedings.

[62] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, Conference Proceedings.

[63] L. Yao, A. Torabi, K. Cho, and N. Ballas, "Describing videos by exploiting temporal structure," in *ICCV*. IEEE, 2015, Conference Proceedings.

[64] V. Ramanathan, J. Huang, S. Abuelhaija, A. Gorban, K. Murphy, and F. F. Li, "Detecting events and key actors in multi-person videos," in *CVPR*. IEEE, 2016, Conference Proceedings.

[65] J. Ba, G. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu, "Using fast weights to attend to the recent past," in *NeurIPS*, 2016, Conference Proceedings.

[66] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *CVPR*. IEEE, 2017, Conference Paper.

[67] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, "Structural-rnn: Deep learning on spatio-temporal graphs," in *CVPR*. IEEE, 2016, Conference Proceedings.

[68] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data." in *AAAI*, 2017, Conference Proceedings.

[69] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[70] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[71] J. Donahue, L. A. Hendricks, S. Guadarrama, and M. Rohrbach, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*. IEEE, 2015, Conference Proceedings.

[72] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv*, 2016.

[73] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning-lecture 6a-overview of mini-batch gradient descent."

[74] M. R. Amer, S. Todorovic, A. Fern, and S.-C. Zhu, "Monte carlo tree search for scheduling activity recognition," in *ICCV*. IEEE, 2013, Conference Proceedings.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, Conference Proceedings.

[76] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich *et al.*, "Going deeper with convolutions," in *CVPR*. IEEE, 2015, Conference Proceedings.

[77] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*. IEEE, 2016, Conference Proceedings, pp. 2818–2826.

[78] M. Qi, Y. Wang, A. Li, and J. Luo, "Sports video captioning by attentive motion representation based hierarchical recurrent neural networks," in *Proceedings of the 1st International Workshop on Multimedia Content Analysis in Sports*, ser. MMSports'18. ACM, 2018, Conference Proceedings.

[79] H. Fan, Z. Xu, L. Zhu, C. Yan, J. Ge, and Y. Yang, "Watching a small portion could be as good as watching all: Towards efficient video classification," in *IJCAI*, 2018, Conference Proceedings.

[80] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[81] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

[82] L. Zhu, Z. Xu, and Y. Yang, "Bidirectional multirate reconstruction for temporal modeling in videos," 2017.

[83] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, Conference Proceedings.

[84] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *ICCV*. IEEE, 2017, Conference Proceedings.

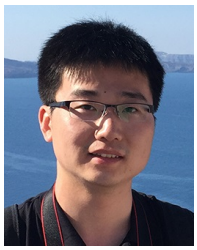
[85] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *CVPR*. IEEE, 2016, Conference Proceedings.

[86] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *International Journal of Computer Vision*, vol. 113, no. 2, pp. 113–127, 2015.

Mengshi Qi received the B.S. and M.S. degree in computer science from Beijing University of Posts and Telecommunications and Beihang University, Beijing, China, in 2012 and 2014, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University. His current research interests include machine learning, computer vision, scene understanding and multimedia retrieval.

Yunhong Wang (M'98-SM'15) received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 1989, and the M.S. and Ph.D. degrees from Nanjing University of Science and Technology, Nanjing, China, in 1995 and 1998, respectively, all in electronics engineering.

She was with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, from 1998 to 2004. Since 2004, she has been a Professor with the School of Computer Science and Engineering, Beihang University, where she is also the Director of Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media. Her research interests include biometrics, pattern recognition, computer vision, data fusion, and image processing.



Jie Qin is currently a research scientist with the Inception Institute of Artificial Intelligence, UAE. He received the B.E. and Ph.D. degrees from Beihang University, China, in 2011 and 2017, respectively. From 2014 to 2015, he was a visiting researcher with The University of Sheffield, UK. From 2017 to 2018, he was a postdoctoral researcher with the Computer Vision Laboratory, ETH Zurich, Switzerland. His current research interests include computer vision and machine learning.

Annan Li received the B.S. and M.S. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2003 and 2006, and the Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He worked in Singapore as a scientist with Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR) and as a postdoctoral research fellow at National University of Singapore, respectively. He currently works at the School of Computer Science and Engineering, Beihang University. His research interests include computer vision, pattern recognition, and statistical learning. He is a member of IEEE.

Jiebo Luo (S'93-M'96-SM'99-F'09) joined the Department of Computer Science at the University of Rochester in 2011, after a prolific career of over 15 years with Kodak Research. He has authored over 400 technical papers and holds over 90 U.S. patents. His research interests include computer vision, machine learning, data mining, social media, and biomedical informatics. He has served as the Program Chair of ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017, and on the Editorial Boards of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON BIG DATA, Pattern Recognition, Machine Vision and Applications, and ACM Transactions on Intelligent Systems and Technology. He is a Fellow of the IEEE, ACM, AAAI, SPIE and IAPR.

Luc Van Gool received the degree in electromechanical engineering at the Katholieke Universiteit Leuven, in 1981. Currently, he is a professor at the Katholieke Universiteit Leuven in Belgium and the ETH in Zurich, Switzerland. He leads computer vision research at both places, where he also teaches computer vision. He has authored more than 200 papers in this field. He has been a program committee member of several major computer vision conferences. His main interests include 3D reconstruction and modeling, object recognition, tracking, and gesture analysis. He received several Best Paper awards. He is a co-founder of 5 spin-off companies.